PR #25690 完整报告

sgl-project/sglang

[Fix] Try to fix error caused by latest cutedsl packages

合并时间：2026-05-19 07:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25690

执行摘要

- 一句话：修复新版 cutdsl 包导致的 CI 错误
- 推荐动作：作为基础设施修复，建议合并。如果团队使用其他 CUDA 版本（如 cu12），需确认此类 extras 标记是否适用。

功能与动机

根据 PR body 关联的 CI 运行日志（https://github.com/sgl-project/sglang/actions/runs/26055697810/job/76604443741）和 vLLM 类似 issue（https://github.com/vllm-project/vllm/pull/40082#issuecomment-4349406309），最新 cutdsl 包的 cu13 extras 会在安装时额外拉取 -libs-base 和 -libs-cu13 两个 wheel，两者写入相同路径但内容不同，导致 GPUModuleOp signature TypeError。

实现拆解

1. 配置依赖改为带 extras 的版本：在 python/pyproject.toml 中，将 flashinfer_python==0.6.11.post1 改为 flashinfer_python[cu13]==0.6.11.post1，将 nvidia-cutlass-dsl==4.5.0 改为 nvidia-cutlass-dsl[cu13]==4.5.0。
2. 调整版本提取正则表达式：在 scripts/ci/cuda/ci_install_dependency.sh 中，修改 flashinfer_python 和 flashinfer_cubin 的版本提取正则，支持可选 extras 部分（如 [cu13]），确保能正确匹配带 extras 的版本字符串。
3. 新增函数 purge_cutlass_libs_base：该函数卸载冲突的 nvidia-cutlass-dsl-libs-base 包，然后从 pyproject.toml 提取 nvidia-cutlass-dsl 的版本，强制重装对应的 nvidia-cutlass-dsl-libs-cu13 包，确保 cu13 风格的绑定文件（如 _gpu_ops_gen.py）覆盖 base 风格的文件。
4. 在安装流程中插入新步骤：在 main() 函数的安装顺序中，在 download_flashinfer_cache 之后、stabilize_flashinfer_jit_paths 之前调用 purge_cutlass_libs_base，确保依赖冲突在后续步骤前被解决。

关键文件：

- scripts/ci/cuda/ci_install_dependency.sh（模块 CI 脚本；类别 infra；类型 infrastructure；符号 purge_cutlass_libs_base, uninstall_stale_flashinfer）：核心变更：新增 purge_cutlass_libs_base 函数来处理 nvidia-cutlass-dsl 的 libs-base 与 libs-cu13 冲突，并在安装流程中调用。同时修改了 flashinfer 版本提取的正则表达式。

- python/pyproject.toml（模块 项目配置；类别 config；类型 configuration）：配置变更：为 flashinfer_python 和 nvidia-cutlass-dsl 添加 [cu13] extras 标记，触发依赖解析拉取 cu13 定制版本。

关键符号：purge_cutlass_libs_base

评论区精华

无 review 评论，讨论仅限 CI 修复本身。

- 暂无高价值评论线程

风险与影响

- 风险：属于纯粹的基础设施修复，不涉及任何源码逻辑变更。风险极低：如果 cu13 extras 在未来版本中被移除，可能需要重新调整；CI 脚本中的 $PIP_UNINSTALL_CMD 和 $PIP_CMD 变量需确保在 CI 环境中正确设置。无安全、性能、兼容性风险。
- 影响：直接修复了因新版 cutdsl 包引入的 CI 失败问题，影响范围仅限于 CI 环境和安装了特定版本 nvidia-cutlass-dsl 的开发环境。用户端无感知，测试流程恢复稳定。
- 风险标记：依赖版本兼容，CI 环境依赖

关联脉络

- 暂无明显关联 PR